

A Journal of Conservation Biogeography

Ensemble distribution models in conservation prioritization: from consensus predictions to consensus reserve networks

Laura Meller^{1,2}*, Mar Cabeza¹, Samuel Pironon², Morgane Barbet-Massin^{3,4}, Luigi Maiorano⁵, Damien Georges² and Wilfried Thuiller²

¹Metapopulation Research Group, Department of Biosciences, P.O. Box 65, 00014 University of Helsinki, Helsinki, Finland, ²Laboratoire d'Ecologie Alpine, UMR-CNRS 5553, Université Joseph Fourier, Grenoble I, BP 53, 38041, Grenoble Cedex 9, France, ³Muséum National d'Histoire Naturelle, UMR 7204 MNHN-CNRS-UPMC, Centre de Recherches sur la Biologie des Populations d'Oiseaux, CP 51, 55 Rue Buffon, 75005 Paris, France, ⁴Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, CT 06520-8106, USA, ⁵Department of Biology and Biotechnologies "Charles Darwin", University of Rome "La Sapienza", Viale dell'Università 32, Rome 00185, Italy

*Correspondence: Laura Meller, Metapopulation Research Group, Department of Biosciences, P.O. Box 65, 00014 University of Helsinki, Helsinki, Finland. E-mail: laura.meller@helsinki.fi

ABSTRACT

Aim Conservation planning exercises increasingly rely on species distributions predicted either from one particular statistical model or, more recently, from an ensemble of models (i.e. ensemble forecasting). However, it has not yet been explored how different ways of summarizing ensemble predictions affect conservation planning outcomes. We evaluate these effects and compare common-place consensus methods, applied before the conservation prioritization phase, to a novel method that applies consensus after reserve selection.

Location Europe.

Methods We used an ensemble of predicted distributions of 146 Western Palaearctic bird species in alternative ways: four different consensus methods, as well as distributions discounted with variability, were used to produce inputs for spatial conservation prioritization. In addition, we developed and tested a novel method, in which we built 100 datasets by sampling the ensemble of predicted distributions, ran a conservation prioritization analysis on each of them and averaged the resulting priority ranks. We evaluated the conservation outcome against three controls: (i) a null control, based on random ranking of cells; (2) the reference solution, based on an expert-refined dataset; and (3) the independent solution, based on an independent dataset.

Results Networks based on predicted distributions were more representative of rare species than randomly selected networks. Alternative methods to summarize ensemble predictions differed in representativeness of resulting reserve networks. Our novel method resulted in better representation of rare species than pre-selection consensus methods.

Main conclusions Retaining information about the variation in the predicted distributions throughout the conservation prioritization seems to provide better results than summarizing the predictions before conservation prioritization. Our results highlight the need to understand and consider model-based uncertainty when using predicted distribution data in conservation prioritization.

Keywords

Consensus predictions, efficiency, optimization, rare species, systematic conservation planning, uncertainty.

INTRODUCTION

Systematic conservation planning (SCP) aims at cost-efficient use of scarce resources available for conservation (Margules & Pressey, 2000). It typically uses optimization tools to identify networks that represent as much biodiversity as possible while keeping the costs or total area to a minimum (Ball *et al.*, 2009; Moilanen *et al.*, 2009). The SCP framework starts by mapping the features of conservation interest within the entire planning region. As such information is rarely

DOI: 10.1111/ddi.12162 http://wileyonlinelibrary.com/journal/ddi

Diversity and Distributions

available and requires tremendous effort to collect, species distribution models (SDMs) have been suggested as providing a useful and cost-efficient tool to generate maps of biodiversity (Guisan & Thuiller, 2005; Wilson *et al.*, 2005; Elith *et al.*, 2006).

Successful use of SDMs in conservation planning requires the model predictions to be accurate and able to identify sites of high suitability for a species (Guisan *et al.*, 2006). At the same time, the model should be able to generalize the prediction outside the locations for which data exist, that is, avoid over-fitting. Both omission and commission errors misdirect allocation of conservation resources and reduce the effectiveness of actions (Rondinini *et al.*, 2006). Validation of models is difficult, as truly independent validation data typically do not exist – and if they did, those data could be used in the planning directly. However, assessment of the uncertainty in predicted distributions, and the impact of this uncertainty on conservation outcomes, is not a standard practice (Langford *et al.*, 2011).

Conservation planning exercises using predicted distribution data have previously relied mostly on predictions from a single statistical technique. However, as the field of species distribution modelling has been evolving rapidly, more reliable alternatives are becoming available. Ensemble modelling, where predictions of species distributions are produced with several statistical techniques together instead of any single modelling technique, has been suggested to improve predictions of the current range of a species (Thuiller, 2004; Araújo et al., 2005; Marmion et al., 2009b) as well as patterns in species richness (Parviainen et al., 2009) and diversity (Mateo et al., 2012). Given that the choice of modelling technique is a major source of variability in the predictions (Pearson et al., 2006; Buisson et al., 2010; Garcia et al., 2011) and that selecting a best model a priori is not straightforward (Elith & Graham, 2009), it seems reasonable to use several, generally well-performing techniques and to thereby assess where their predictions agree or disagree. If different statistical techniques fit similar responses of species occurrence to environmental factors, then these techniques would be expected to make similar predictions. Combining such techniques within an ensemble would add little additional information compared with using just one single technique. However, statistical techniques often differ in how they are affected by geographical range properties (Marmion et al., 2009a) and sensitivity to sample size and species prevalence (Wisz et al., 2008). Therefore, variability across techniques can provide an estimate of an important source of uncertainty. Using that information in reserve selection might be expected to yield more reliable identification of conservation priorities.

To date, ensemble modelling has been applied to conservation planning in order to assess the impacts of climate change on current areas of conservation priority (Coetzee *et al.*, 2009; Araújo *et al.*, 2011; Kujala *et al.*, 2011). It has also been used to identify protected area networks that are resilient to climate change (Carroll *et al.*, 2010) and robust

against uncertainty in the predictions (Carvalho *et al.*, 2011). However, the consequences of subjective decisions in using the ensemble outputs in conservation planning have not yet been examined. As ensembles are collections of predictions, there is a preference towards summarizing the information across the multiple projections to aid decision-making.

However, there are several possible ways of summarizing the ensemble of predictions into one, or a few, predictive map(s) of species distribution (Araújo & New, 2007; Marmion *et al.*, 2009b). In previous studies, ensembles have been summarized *a priori* to spatial conservation prioritization so that the summary maps are used as inputs for identifying conservation priorities. Alternatively, the range of predictions in the full ensemble could be used to identify multiple sets of conservation priorities, and a summary could be made of those priorities *a posteriori*. This alternative approach is expected to retain more information about the variability across model predictions throughout the conservation prioritization exercise while demanding less computation power than running conservation prioritization across all possible combinations of the full ensemble.

Here, we evaluate and compare the pre-selection and post-selection consensus approaches in conservation prioritization. As the latter approach has not been used in any previous study, we develop and assess a novel parsimonious method for using data representative of the full ensemble in the conservation prioritization process. We evaluate the performance of each approach by comparing the representativeness of the resulting protected area network with that achieved by (1) randomly selected networks, to test for overall performance, (2) a quasi-independent reference dataset, to test for solution efficiency and (3) a truly independent evaluation dataset. Essentially, we ask: Would conservation decisions based on ensemble predictions of species distributions be different depending on how the ensemble is pre-processed? We discuss how the choice of approach may depend on the properties of species data that are available, and why it is critical to assess the reliability of species distribution modelling in any given conservation planning context.

METHODS

Bird distribution data

We used extent of occurrence data of 158 bird species in the Western Palaearctic region to conduct the analysis. These species are in the Annex I of the European Union Birds Directive and thereby have a legal conservation status at the EU level. Distribution maps from The Birds of the Western Palaearctic database (BWPi, 2006) were digitized at a resolution of 50 \times 50 km (Barbet-Massin *et al.*, 2012b). We calibrated the models using species distributions across the whole Western Palaearctic region to cover the full extent of their niches (Barbet-Massin *et al.*, 2010).

Climate and land use data

Current climate was represented by 19 bioclimatic variables from the WorldClim database (Hijmans *et al.*, 2005) at 50-km grid resolution for calibrating the models and 10' resolution for projecting them. We pre-selected variables based on a principal components analysis and a cross-correlation matrix to avoid retaining highly correlated variables. We finally kept five climatic variables in the models: temperature seasonality (standard deviation *100), maximum temperature of the warmest month, minimum temperature of the coldest month, precipitation of the wettest month and precipitation of the driest month.

Current land cover for the whole Palaearctic was represented by GlobCover 2009 (Arino *et al.*, 2012) at 300-m resolution. We upscaled the data to 50×50 km and 10' resolution by calculating the area of each land cover type within each pixel, using the level 1 classification (i.e. built-up areas, arable lands, permanent crops, grasslands, forests and others). Because the built-up and arable lands variables were highly correlated, we removed built-up areas from the set of variables. In addition, we estimated the Simpson diversity index using the proportional cover of each land cover class as a weight. This can be interpreted as a measure of land use heterogeneity.

Species distribution modelling

Species distribution models were calibrated over the whole Palaearctic region at a resolution of 50×50 km and then projected at 10' resolution. We used five different algorithms available within the BIOMOD framework (Thuiller et al., 2009) to obtain an ensemble of predicted distributions: generalized additive models (GAM), boosting regression trees (BRT), classification tree analysis (CTA), multiple adaptive regression splines (MARS) and random forest (RF). We chose a restricted set of algorithms because it was beyond the aim of the study to cover the complete range of available algorithms, and these have been demonstrated to be performing well in species distribution modelling across a range of scales and situations (Thuiller et al., 2003; Elith et al., 2006; Lawler et al., 2006; Prasad et al., 2006). To calibrate the models, and to evaluate their performance through cross-validation, we partitioned the original data into two subsamples: one for calibration-validation and the other for evaluation (Friedman et al., 2009). For the calibration-validation subset, a fivefold internal cross-validation was carried out by randomly sampling 62.5% of the calibration data and testing the predictions against the remaining 37.5% of the data. The resulting projections were then validated with the 20% of data that were not used for model calibration or internal evaluation, using the true skill statistic (TSS) as an evaluation metric (Allouche et al., 2006). With five repetitions and five modelling algorithms, we obtained an ensemble of 25 predicted distributions for each of our species (Fig. 1).

To explore whether the performance of different reserve selection approaches varies for different types of species, we grouped species based on their model-related uncertainty using range size and the median correlation (Pearson's R) between predictions across models (i.e. a measure of how different the predictions are from the different models and the repetitions). We calculated a pairwise Euclidean distance matrix between species based on these two parameters and ran a hierarchical cluster analysis on this matrix (Venables & Ripley, 2003). The number of clusters was determined by maximizing the Silhouettes index (Rousseeuw, 1987). This resulted in three distinguishable classes, within which cluster 1 contained 84 species whose cross-model correlation and range size were both lower than for species in the other two clusters (Fig. S2; Table S1a). In addition to presenting results for all species considered in the analysis, we also present results separately for species in cluster 1, as the high variability exhibited by these species within the model ensemble makes them a good test of ability of different consensus methods to inform conservation planning. These species are also rare (occurring in <20% of the cells in the study area), which make them of conservation interest. We refer to the species in cluster 1 as 'rare species' in the Results section.

Ensemble prediction datasets

To obtain the datasets for the 'pre-selection consensus technique', we summarized the ensemble of predicted species distributions using four different consensus techniques available within BIOMOD: (1) committee averaging (MeanTSS), where probabilities of occurrence from different models were first transformed to presences and absences with a threshold that maximizes the value of the true skill statistic (Allouche et al., 2006) and then averaged (Thuiller et al., 2009); (2) weighted mean probability (MeanWMP), where a weight based on the evaluation scores was first assigned to the probabilities and no transformation to presences and absences occurs (Marmion et al., 2009b; Thuiller et al., 2009); (3) binary committee averaging (BinTSS), where the committee average probabilities were transformed back to presences and absences with the threshold that maximized the TSS score during the cross-validation procedure; (4) binary weighted mean probability (BinWMP), where the weighted mean probabilities were transformed to presences and absences with a threshold that maximizes the TSS of the ensemble predictions (Thuiller et al., 2009).

We produced raster grids of the standard deviation across the ensemble for each species and grid cell. The standard deviation for each cell, together with the weighted mean probability, was used in an approach called distribution discounting (DistrDisc; Moilanen *et al.*, 2006), previously implemented in the conservation planning software Zonation (see below).

Additionally, we produced datasets for the 'post-selection consensus technique' by randomly sampling one probability value of the 25 values available for each species in each grid cell and repeated this sampling 100 times, thereby achieving 100 datasets altogether.



Figure 1 We used two evaluation datasets in the analysis: a reference dataset where species distribution data for the analysis were processed in an alternative way, to provide higher quality distributional information, and an independent dataset from another data source. The pre-selection consensus datasets were obtained by summarizing the ensemble predictions in four different ways. The distribution discounting dataset had pre-selection consensus of predicted distributions combined with layers of standard deviations across the ensemble predictions separately for each species. The post-selection consensus data contained 100 datasets where a probability of occurrence was randomly sampled from the full ensemble for each species in each cell. The conservation prioritization for the reference, independent and pre-selection consensus datasets was based on the layers of (predicted) species distributions. For the distribution discounting set, the standard deviation was subtracted from the probability of occurrence. Conservation priority ranking was generated for each of the 100 post-selection consensus datasets separately, after which the mean rank of each cell was calculated to produce the final post-selection consensus ranking.

We excluded species that: (1) were missing from the evaluation datasets or had a different taxonomic interpretations; or (2) for which the true skill statistics validation score was lower than 0.3 for one or more models, meaning that the model fit was poor and predictions unreliable (Araújo *et al.*, 2011; see Fig. S1a and Table S1a for details on the validation statistics). We retained 146 species in the final analyses.

Evaluation datasets

Reference evaluation dataset

We used reference data for the species in our analysis (Maiorano *et al.*, 2013) to assess how closely the reserve network identified with predicted distributions matched a

network derived using higher quality data. In the reference dataset, the extent of occurrence (EOO) maps in the Birds of the Western Palaearctic database (BWPi, 2006) were complemented with data from The EBCC Atlas of European Breeding Birds (Hagemeijer & Blair, 1997). EOO maps were filtered with land cover and elevation data with resolution of 0.9 km², following published literature and expert opinion, considering three main environmental variables: land use based on GlobCover 2009 (Arino et al., 2012), elevation and distance to water (see Maiorano et al., 2013 for more details). Land cover was classified according to suitability for each species. Presences were assigned to cells that contained primary habitat for a species within its EOO range. The resulting ranges were aggregated to presences and absences at 10' resolution so that a species was considered present in a cell if there was at least one suitable 0.9-km² cell within a 10' cell.

The agreement between the original data used for modelling and the reference evaluation data, quantified as TSS scores, varied between 0.21 and 0.94, with a median of 0.81. The TSS scores for the agreement between predicted distributions and reference evaluation data varied between -0.03and 0.97 (see Fig. S1b and Table S1b for details).

Independent evaluation dataset

We used another evaluation dataset (BirdLife International & NatureServe, 2012) to test the performance of our approaches against data from independent sources. The independent evaluation dataset includes distribution information compiled from a variety of sources, including specimen localities from BirdLife's Point Locality Database and the Global Biodiversity Information Facility (GBIF), observer records, documented occurrences in Important Bird Areas, distribution atlases derived from systematic surveys, distribution maps in field guides and other handbooks and expert opinion. The agreement between the original data used for modelling and the independent evaluation data, quantified as TSS scores, varied between 0 and 0.98, with a median of 0.77. The TSS scores for the agreement between predicted distributions and independent evaluation data varied between -0.31 and 0.99. The agreement between reference data and independent data varied between 0.04 and 1.00, with a median of 0.71 (see Fig. S1c and Table S1c for details).

Reserve selection

To identify priority areas for conservation, we used the Zonation v.3.0 software for spatial conservation prioritization (Conservation Biology Informatics Group, University of Helsinki, Helsinki, Finland; Moilanen *et al.*, 2012). Zonation identifies areas that are important for retaining suitable habitat simultaneously for all the species included in the analysis, to provide decision support and facilitate cost-efficient conservation decisions. The output of Zonation is a hierarchical map of the landscape, based on the biodiversity value of the sites. The software operates through backwards-iterative heu-

ristics, at each step calculating conservation value for each site and removing the one with the lowest conservation value. We used the algorithm called core-area Zonation that calculates the conservation value δ_i for each site as

$$\delta_i = \max_j (q_{ij} w_j / c_i)$$

where q_{ij} is the proportion of the distribution of species *j* located in site *i* among the sites that are remaining in the landscape, w_j is the species-specific weight for species *j*, and c_i is the cost of site *i*. As we weighted all species equally and did not use information about land cost in our analysis, the conservation value was purely determined by the species that had the largest proportion of its remaining distribution in cell *i*.

To assess the performance of the ensemble prediction datasets in reserve selection, we produced three different controls (Fig. 1): (1) a null control, based on 100 random rankings of cells disregarding any information about the species; (2) the reference solution, with conservation priority ranking based on the reference evaluation dataset; and (3) the independent solution, with conservation priority ranking based on the independent evaluation dataset.

We derived a conservation priority ranking with each of the four 'pre-selection consensus' datasets (Fig. 1). We also generated a conservation priority ranking that considered uncertainty through the distribution discounting analysis within Zonation (Moilanen et al., 2006, 2012) using the MeanWMP dataset in combination with the standard deviation across the ensemble for each species (DistrDisc). Here, the standard deviation of each species and grid cell is subtracted from the probability of occurrence for that species and cell. Finally, to obtain the 'post-selection consensus' reserve network, we first derived a conservation priority ranking for each of the 100 datasets we had sampled across the full ensemble. We calculated the mean rank for each cell across these rankings and reranked the cells by the mean rank. The resulting ranking was our 'post-selection consensus' prioritization (PostCons). The number of sampled datasets should be balanced so that the samples are representative of the ensemble while the analysis is not too heavy in terms of computation. In our case, the Spearman's correlation r_s between average rank across 10 and 100 runs was already higher than 0.95 (see Fig. S3).

Comparative analysis: similarity and representativeness

We explored the similarity between networks by quantifying: (1) pairwise spatial overlaps between the highest 10% priorities of ensemble prediction versus reference and independent solutions; and (2) pairwise correlations between the overall priority rankings of ensemble prediction and reference and independent solutions. To compare the performance of our methods with the null control, we quantified the number of times each species was better represented in the ensemble prediction-based networks than in the networks based on random ranking of cells. We quantified representativeness of species, according to data in the evaluation datasets, in the reserve networks based on the ensemble prediction datasets. We used pairwise Wilcoxon's signed-rank tests to determine whether species are consistently better represented in one network or another. The R script and Zonation set-up for computing the analyses are provided in Appendix S1 in Supporting Information.

RESULTS

Spatial similarity in conservation priorities

The conservation priorities obtained with the post-selection consensus approach were most similar to both the reference and independent priorities, with 39% and 45% spatial overlap of the top 10% fraction for the reference and independent networks, respectively, and Spearman's correlation r_s of 0.551 and 0.599 for the overall ranking of cells (Fig. 2; Table 1.). The overlap between pre-selection consensus networks and the reference network ranged between 21% and 31% and Spearman's correlation r_s between 0.385 and 0.467, whereas the distribution discounting network had 28% cells in common with the reference network, and the Spearman's correlation r_s was 0.411. Overlap between pre-selection consensus networks and independent network ranged between 27% and 33% and was 29% between distribution discounting network and independent network, while the respective Spearman's correlations r_s were 0.409–0.499 and 0.468.

Representativeness

Both the evaluation networks (reference and independent) as well as the ensemble prediction networks were consistently more representative than random networks for 43-59% of all species and 67-88% of the rare species, when evaluated with the reference evaluation data, and for 33-55% of all species and 54-83% of rare species, when evaluated with independent evaluation data (Table 2). However, the differences in representation between random networks and the network based on MeanTSS were not statistically significant for all species against either reference or independent data (Wilcoxon's signed-rank test; median P = 0.178 in pairwise comparisons between MeanTSS and the 100 random networks against reference data and median P = 0.491 against independent data), or for all species in the post-selection consensus network when evaluated against independent data (median P = 0.086). In all other cases, the differences were statistically significant (Wilcoxon's signed-rank test; P < 0.05).

Both evaluation networks represented significantly larger proportions of species ranges than any of the networks obtained with the predicted distributions, both for all species and for rare species (Fig. 3; Tables 3 and S1).

Post-selection consensus yielded consistently higher representation for rare species than any of the pre-selection consensus networks, or the distribution discounting network, when evaluated against the reference data (Fig. 3b; Table 3). These differences were statistically significant (Wilcoxon's signed-rank test; P < 0.05) in all cases except between post-selection consensus and BinWMP (P = 0.076). When evaluated against the independent evaluation data, the PostCons network was more representative than any of the other ensemble prediction networks for rare species (Fig. 3d; Table 3). There were no significant differences between post-selection consensus and pre-selection consensus or distribution discounting for all species, neither against reference nor independent evaluation data (Table 3).

DISCUSSION

We have presented a new approach to using ensemble model outputs in reserve selection: the post-selection consensus approach, where the full range of predictions is sampled to provide input for spatial conservation prioritization. In our study, the post-selection consensus approach resulted in consistently better conservation outcomes for rare species than using a pre-selection consensus summary of the model predictions.

Selecting an appropriate approach based on prediction variability

Patterns of similarity in model predictions across the ensemble can provide valuable information and guide the choice of approach for conservation prioritization. In our analysis, widespread species generally had high correlation between model predictions (mean Pearson's R 0.94; 1. quartile 0.92; 3. quartile 0.95), and no consistently significant differences were found between the performance of pre- and post-selection consensus techniques when such species were included in the comparison. Species that were present in <20% of the cells in the study area exhibited higher variation across models (mean Pearson's R 0.82; 1. quartile 0.77; 3. quartile 0.89), and for these species, the post-selection consensus approach was consistently better than the pre-selection consensus techniques. However, species with very few occurrences were often represented poorly in the ensemble prediction networks, particularly when they also had exceptionally low cross-model correlation (Pearson's R < 0.75). This suggests that their distributions were not described by model predictions accurately enough. These species occurred in our study area only at the margins of their ranges or had a scattered range, which are probable explanations for the high variability. Such species can be included in the planning exercise by directly using the available observation data instead of modelled distributions, while using predicted distributions for the rest of the species (Carvalho et al., 2010).

Virtual experiments can address limitations of this study

A potential source of uncertainty, not addressed in our study, arises from our choice of the reference evaluation data



Figure 2 Nested top fractions of conservation priority rankings for: (a) reference, (b) independent, (c) BinTSS, (d) BinWMP, (e) MeanTSS, (f) MeanWMP, (g) distribution discounting and (h) post-selection consensus.

		Pre-selecti	on consensus					
		BinTSS	BinWMP	MeanTSS	MeanWMP	DistrDisc	Post-selection consensus	
Reference	Overlap (%)	21	27	32	31	28	39	
	Correlation (r_s)	0.385	0.422	0.398	0.467	0.411	0.551	
Independent	Overlap (%)	27	27	29	33	29	45	
	Correlation (r_s)	0.487	0.467	0.409	0.499	0.458	0.599	

Table 1 Spatial overlap between evaluation and test conservation priorities (defined as best 10% of the priority ranking) as well as the Spearman's correlation r_s of the priority rankings.

Table 2 The number and proportion of species that were consistently (more than 95 times of 100) represented better in evaluation and ensemble conservation priority networks than in randomly selected networks of equal size, evaluated with (a) reference and (b) independent data. The values are reported separately for rare species (n = 84) and all species (n = 146).

	Reference	Pre-selection	n consensus					
		BinTSS	BinWMP	MeanTSS	MeanWMP	DistrDisc	Post-selection consensus	
(a)								
Rare species N	74	61	67	56	65	68	62	
%	88.1	72.6	79.8	66.7	77.4	81	73.8	
All species N	86	72	79	62	73	76	70	
%	58.9	49.3	54.1	42.5	50	52.1	47.9	
		Pre-select	ion consensus					
	Independent	BinTSS	BinWMP	MeanTSS	MeanWMP	DistrDisc	Post-selection consensus	
(b)								
Rare species N	70	63	63	45	63	67	64	
%	83.3	75	75	53.6	75	79.8	76.2	
All species N	80	67	71	48	68	73	69	
%	54.8	45.9	48.6	32.9	46.6	50	47.3	

that were based on an expert assessment of habitat suitability within known extent of occurrence, rather than on direct surveys or observations of species occurrence. We further assessed pre- and post-selection approaches against an independent evaluation dataset, which is also likely to contain errors. However, the differences in performance of consensus methods were consistent in both assessments.

Using empirical, and therefore uncertain, data in our study means that the true performance of the approaches cannot be known. This shortcoming could be addressed in the future through virtual experiments (Langford et al., 2009; Zurell et al., 2010). However, an empirical study such as ours provides a good starting point for identifying relevant questions that virtual studies can better address. Relevant questions arising from our study are: (1) What role does variability in ensembles play in the performance of different consensus methods? (2) Does community nestedness affect the performance of pre- versus post-selection consensus? and (3) How robust are different approaches to increased uncertainty, for example, under climate change? The emphasis on robustness should become even more pronounced when distributions are projected into the future, and several socio-economic scenarios and climatic circulation models may be involved in addition to the statistical techniques for SDMs (Buisson *et al.*, 2010; Garcia *et al.*, 2011).

We compared the performance of different approaches with a network of fixed size. However, it is worth noting that increasing the coverage of protected area network reduces sensitivity to errors, as well as increasing the likelihood of population persistence (Cabeza & Moilanen, 2003; Hannah *et al.*, 2007).

Using SDM predictions in conservation planning needs careful consideration

We have shown that conservation decisions based on predicted species distributions are sensitive to the approach to summarizing the output of ensembles. Our proposed method (post-selection consensus) provided a better conservation outcome than pre-selection consensus methods, which summarize the model ensemble before conservation prioritization. However, there were notable differences between the pre-selection consensus techniques. For example, the network based on weighted mean probabilities of occurrence transformed to binary presences and absences was not significantly less representative of rare species than the



Figure 3 Representativeness of the reserve networks, as defined by proportions of species ranges, retained in the 10% of cells with highest priority rank, and evaluated against the reference evaluation data for (a) all and (b) rare species, as well as the independent evaluation data for (c) all and (d) rare species. The box plots display the median as well as the 50% (box) and 95% (whiskers) confidence intervals, and the dashed horizontal line indicates the median representation in the evaluation data.

post-consensus network, when evaluated against reference data. The results of this first case study highlight the need to carefully contemplate the use of SDM outputs in spatial conservation prioritization.

We calibrated our SDMs with comprehensive information about species distributions in the study area. In most realworld situations, available data would be less comprehensive and would therefore exhibit higher levels of error and bias. It is critical to first assess whether data are appropriate for SDM before proceeding down this path. Data error, uncertainty and model reliability have attracted considerable research interest over recent years (Reddy & Dávalos, 2003; Graham *et al.*, 2004; Barry & Elith, 2006; Barbet-Massin *et al.*, 2012a), and all of this understanding applies to both pre- and post-consensus approaches in conservation prioritization. Neither pre- nor post-consensus reserve selection are reliable if error and bias in the original data render model predictions unreliable.

Using high-quality reference data on species distributions provided the best outcome in our study. This serves as a reminder that, when available, representative and unbiased datasets can serve as a suitable basis for conservation planning even without modelling (Gaston & Rodrigues, 2003). Any conservation analysis should start by assessing whether **Table 3** Performance of test conservation priority networks with respect to representativeness. The sums and *P*-values of Wilcoxon's signed-rank tests comparing (a) evaluation networks to ensemble prediction networks and (b) post-selection consensus network to all other networks using the reference and independent evaluation data.

				Pre-selection consensus							
				BinTSS	BinWM	IP N	MeanTSS		nWMP	DistrDisc	Post-selection consensus
(a)											
Rare species	Reference	e Sum of	Sum of pos. ranks		3216.5	3	3235	3419		3462	2784
		Sum of	neg. ranks	203	353.5		335	151		108	786
		P		<0.00	01 <0.00)1	< 0.001	<0	.001	< 0.001	< 0.001
	Independ	lent Sum of	pos. ranks	3260.5	3251.5	3	3137	3272		3350	3186.5
		Sum of	neg. ranks	309.5	318.5		349	214		220	383.5
		P		<0.00	01 <0.00)1	< 0.001	<0	.001	< 0.001	< 0.001
All species	Reference	e Sum of	pos. ranks	8106.5	7993	8	3465	8712	.5	8783	8522.5
		Sum of	neg. ranks	2333.5	2738	2	2266	1872	.5	1802	2062.5
		Р		< 0.00	01 <0.00)1	< 0.001	<0	.001	< 0.001	< 0.001
	Independ	lent Sum of	pos. ranks	7608.5	7824.5	7	7881.5	7963		8044.5	9742
		Sum of	neg. ranks	2976.5	2906.5	2	2558.5	2477		2686.5	989
		Р		<0.00	0.00)1	< 0.001	<0	.001	< 0.001	< 0.001
				Pre-selection consensus					ensus		
					Evaluation	BinTS	S Bi	inWMP	MeanTSS	MeanWMI	- P DistrDisc
(b)											
PostCons		Rare species	Sum of pos	ranks	786	2638	2	184	2651.5	2689.5	2712
(Reference)			Sum of neg	. ranks	2784	848	13	386	918.5	880.5	858
()			Р		< 0.001	<0.0	001	0.076	< 0.001	< 0.001	< 0.001
		All species	Sum of pos	. ranks	2062.5	5209	44	4455.5	5603	5567.5	5703
		1	Sum of neg	. ranks	8522.5	5376	62	275.5	4837	5163.5	5028
			Р		< 0.001	0.8	37	0.076	0.446	0.694	0.51
PostCons (Independent)		Rare species	Sum of pos	. ranks	772	2591	21	166	2653	2664	2689
			Sum of neg	. ranks	2798	895	14	404	917	906	881
			P		< 0.001	0.0	002	0.023	< 0.001	0.001	0.001
		All species	Sum of pos	. ranks	989	4888	45	591.5	6018	4959	5021.5
		ĩ	Sum of neg	. ranks	9742	5697	58	348.5	4422	5626	5563.5
			P		< 0.001	0.4	425	0.21	0.112	0.511	0.593

models are needed for that analysis, and whether the data are suitable for SDM. In our study, we further showed that very rare species are particularly difficult to model accurately, and this variability may not be apparent from the TSS evaluation scores. Such species may be inadequately represented in the reserve networks because of incorrect model predictions. We caution against disillusion with respect to statistical techniques and call for using best practices identified by SDM research (Barry & Elith, 2006; Elith *et al.*, 2006; Guisan *et al.*, 2006) in conservation planning as well as other fields of application.

ACKNOWLEDGEMENTS

We are grateful to Simon Ferrier, Bill Langford and an anonymous referee for their constructive comments on this manuscript. We thank BirdLife International and Nature-Serve for providing their data and acknowledge their endeavour for such a database. The research leading to these results received funding from the European Research Council under the European Community's Seventh Framework Programme FP7/2007-2013 project TEEMBIO (Grant Agreement No. 281422). Species distributions were modelled using the CIMENT infrastructure (https://ciment. ujf-grenoble.fr), supported by the Rhône-Alpes region (GRANT CPER07_13 CIRA: www.ci-ra.org). LMeller acknowledges LUOVA Graduate School for funding. MC and LMeller acknowledge funding from EU FP7 project RESPONSES (Grant Agreement No. 244092). MBM acknowledges funding from EU FP7-PEOPLE-2011-IOF project BIRDCHANGE.

REFERENCES

Allouche, O., Tsoar, A. & Kadmon, R. (2006) Assessing the accuracy of species distribution models : prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232.

- Araújo, M.B. & New, M. (2007) Ensemble forecasting of species distributions. *Trends in ecology & evolution*, **22**, 42–47.
- Araújo, M.B., Whittaker, R.J., Ladle, R.J. & Erhard, M. (2005) Reducing uncertainty in projections of extinction risk from climate change. *Global Ecology and Biogeography*, **14**, 529–538.
- Araújo, M.B., Alagador, D., Cabeza, M., Nogués-Bravo, D. & Thuiller, W. (2011) Climate change threatens European conservation areas. *Ecology Letters*, 14, 484–492.
- Arino, O., Perez Ramos, J.J., Kalogirou, V., Bontemps, S., Defourny, P. & Van Bogaert, E. (2012) *Global land cover map for 2009 (GlobCover 2009)*, European Space Agency (ESA) & Université catholique de Louvain (UCL), Frascati.
- Ball, I.R., Possingham, H.P. & Watts, M.E. 2009) Marxan and relatives: software for spatial conservation prioritization. Spatial conservation prioritization. quantitative methods & computational tools (ed. by A. Moilanen, K.A. Wilson and H.P. Possingham), pp. 185–195. Oxford University Press, New York.
- Barbet-Massin, M., Thuiller, W. & Jiguet, F. (2010) How much do we overestimate future local extinction rates when restricting the range of occurrence data in climate suitability models? *Ecography*, **33**, 878–886.
- Barbet-Massin, M., Jiguet, F., Albert, C.H. & Thuiller, W. (2012a) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, **3**, 327–338.
- Barbet-Massin, M., Thuiller, W. & Jiguet, F. (2012b) The fate of European breeding birds under climate, land-use and dispersal scenarios. *Global Change Biology*, **18**, 881–890.
- Barry, S. & Elith, J. (2006) Error and uncertainty in habitat models. *Journal of Applied Ecology*, **43**, 413–423.
- BirdLife International and NatureServe. (2012) Bird species distribution maps of the world. Version 2.0.
- Buisson, L., Thuiller, W., Casajus, N., Lek, S. & Grenouillet, G. (2010) Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, **16**, 1145–1157.
- BWPi, (2006) *Birds of the Western Palearctic interactive*. Bird-Guides, Oxford, UK.
- Cabeza, M. & Moilanen, A. (2003) Site-selection algorithms and habitat loss. *Conservation Biology*, **17**, 1402–1413.
- Carroll, C., Dunk, J.R. & Moilanen, A. (2010) Optimizing resiliency of reserve networks to climate change: multispecies conservation planning in the Pacific Northwest, USA. *Global Change Biology*, **16**, 891–904.
- Carvalho, S.B., Brito, J.C., Pressey, R.L., Crespo, E. & Possingham, H.P. (2010) Simulating the effects of using different types of species distribution data in reserve selection. *Biological Conservation*, **143**, 426–438.
- Carvalho, S.B., Brito, J.C., Crespo, E.G., Watts, M.E. & Possingham, H.P. (2011) Conservation planning under climate change: toward accounting for uncertainty in predicted species distributions to increase confidence in conservation investments in space and time. *Biological Conservation*, **144**, 2020–2030.

- Coetzee, B.W.T., Robertson, M.P., Erasmus, B.F.N., Van Rensburg, B.J. & Thuiller, W. (2009) Ensemble models predict Important Bird Areas in southern Africa will become less effective for conserving endemic birds under climate change. *Global Ecology and Biogeography*, **18**, 701–710.
- Elith, J. & Graham, C.H. (2009) Do they? how do they? why do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, **32**, 66–77.
- Elith, J., Graham, C.H., Anderson, R.P. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.
- Friedman, J., Hastie, T. & Tibshirani, R. (2009) *Elements of statistical learning: data mining, inference, and prediction.* Springer-Verlag, New York.
- Garcia, R.A., Burgess, N.D., Cabeza, M., Rahbek, C. & Araújo, M.B. (2011) Exploring consensus in 21st century projections of climatically suitable areas for African vertebrates. *Global Change Biology*, 18, 1253–1269.
- Gaston, K.J. & Rodrigues, A.S.L. (2003) Reserve selection in regions with poor biological data. *Conservation Biology*, 17, 188–195.
- Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in ecology & evolution*, **19**, 497–503.
- Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.
- Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J.M.C., Aspinall, R. & Hastie, T. (2006) Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology*, **43**, 386–392.
- Hagemeijer, E.J.M. & Blair, M.J. (eds.) (1997) The EBCC Atlas of European breeding birds: their distribution and abundance. T & A.D, Poyser, London.
- Hannah, L., Midgley, G., Andelman, S., Araújo, M.B., Hughes, G., Martinez-Meyer, E., Pearson, R. & Williams, P. (2007) Protected area needs in a changing climate. *Frontiers in Ecology and the Environment*, 5, 131–138.
- Hijmans, R. J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25, 1965–1978.
- Kujala, H., Araújo, M.B., Thuiller, W. & Cabeza, M. (2011) Misleading results from conventional gap analysis – Messages from the warming north. *Biological Conservation*, 144, 2450–2458.
- Langford, W.T., Gordon, A. & Bastin, L. (2009) When do conservation planning methods deliver? Quantifying the consequences of uncertainty. *Ecological Informatics*, **4**, 123–135.
- Langford, W.T., Gordon, A., Bastin, L., Bekessy, S.A., White, M.D. & Newell, G. (2011) Raising the bar for systematic conservation planning. *Trends in ecology & evolution*, **26**, 634–640.

- Lawler, J.J., White, D., Neilson, R.P. & Blaustein, A.R. (2006) Predicting climate-induced range shifts: model differences and model reliability. *Global Change Biology*, **12**, 1568– 1584.
- Maiorano, L., Amori, G., Capula, M., Falcucci, A., Masi, M., Montemaggiori, A., Pottier, J., Psomas, A., Rondinini, C., Russo, D., Zimmermann, N.E., Boitani, L. & Guisan, A. (2013) Threats from climate change to terrestrial vertebrates hotspots in Europe. *PLoS One*, 8, e74989.
- Margules, C.R. & Pressey, R.L. (2000) Systematic conservation planning. *Nature*, **405**, 243–253.
- Marmion, M., Luoto, M., Heikkinen, R.K. & Thuiller, W. (2009a) The performance of state-of-the-art modelling techniques depends on geographical distribution of species. *Ecological Modelling*, **220**, 3512–3520.
- Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R.K. & Thuiller, W. (2009b) Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions*, **15**, 59–69.
- Mateo, R.G., Felicísimo, Á.M., Pottier, J., Guisan, A. & Muñoz, J. (2012) Do stacked species distribution models reflect altitudinal diversity patterns? *PLoS One*, **7**, e32586.
- Moilanen, A., Wintle, B.A., Elith, J. & Burgman, M. (2006) Uncertainty analysis for regional-scale reserve selection. *Conservation Biology*, **20**, 1688–1697.
- Moilanen, A., Kujala, H. & Leathwick, J.R. (2009) The Zonation framework and software for conservation prioritization. Spatial conservation prioritization. quantitative methods & computational tools (ed. by A. Moilanen, K.A. Wilson and H.P. Possingham), pp. 196–210. Oxford University Press, New York.
- Moilanen, A., Meller, L., Leppänen, J., Montesino Pouzols, F., Arponen, A. & Kujala, H. (2012) Spatial conservation planning framework and software Zonation. Version 3.1 User manual, Helsinki.
- Parviainen, M., Marmion, M., Luoto, M., Thuiller, W. & Heikkinen, R.K. (2009) Using summed individual species models and state-of-the-art modelling techniques to identify threatened plant species hotspots. *Biological Conservation*, **142**, 2501–2509.
- Pearson, R.G., Thuiller, W., Araújo, M.B., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T.P. & Lees, D.C. (2006) Model-based uncertainty in species range prediction. *Journal of Biogeography*, 33, 1704– 1711.
- Prasad, A.M., Iverson, L.R. & Liaw, A. (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, **9**, 181–199.
- Reddy, S. & Dávalos, L.M. (2003) Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, **30**, 1719–1727.
- Rondinini, C., Wilson, K.A., Boitani, L., Grantham, H. & Possingham, H.P. (2006) Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecology Letters*, **9**, 1136–1145.

- Rousseeuw, P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.
- Thuiller, W. (2004) Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, **10**, 2020–2027.
- Thuiller, W., Araújo, M.B. & Lavorel, S. (2003) Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science*, **14**, 669–680.
- Thuiller, W., Lafourcade, B., Engler, R. & Araújo, M.B. (2009) BIOMOD a platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369–373.
- Venables, W.N. & Ripley, B.D. (2003) *Modern applied statistics with S.* Springer-Verlag, New York.
- Wilson, K.A., Westphal, M.I., Possingham, H.P. & Elith, J. (2005) Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation*, **122**, 99–112.
- Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H. & Guisan, A. (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14**, 763–773.
- Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Münkemüller, T., Nehrbass, N., Pagel, J., Reineking, B., Schröder, B. & Grimm, V. (2010) The virtual ecologist approach: simulating data and observers. *Oikos*, **119**, 622–635.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Script and other set-up files for running the analysis.

Table S1 Species prevalences, validation statistics and repre-sentation in conservation priority networks.

Figure S1 True skill statistics of the model ensemble and agreement between datasets.

Figure S2 Species clusters based on range size and cross-model correlation.

Figure S3 Correlations between post-selection consensus rankings with different numbers of datasets.

BIOSKETCH

Laura Meller (MSc, University of Helsinki) is interested in uncertainty related to conservation planning at different levels, ranging from the distribution of biodiversity to societal responses to environmental change. In the course of her PhD studies, her aim is to assess and develop tools to fill gaps in our understanding of present conservation needs and uncertainty about future changes.

Ensemble distribution models in conservation prioritization

Author contributions: L. Meller and W.T. conceived the ideas; L. Meller, M.C. and W.T. designed the analysis; M.B-M. and L. Maiorano compiled data for the analysis; S.P., D.G. and W.T. produced the ensemble predictions of species distributions; L. Maiorano proceeded the refined bird

distributions. L. Meller executed the analyses and L. Meller led the writing with contributions from all co-authors.

Editor: Simon Ferrier